

Using Ensemble AI Models to Better Predict Ramping Events

Craig Collier

Chief Meteorologist, Head of Operations
Energy Forecasting Solutions

ESIG Forecasting and Markets Workshop
Denver, CO | June 15, 2026



Why AI Ensembles for Ramp Prediction?



The Challenge

Wind power ramp events drive disproportionate economic exposure for grid operators and forecast vendors. System imbalance costs significantly higher during ramp hours.

The Opportunity

AI weather models (GraphCast, Pangu-Weather, FourCastNet) have matched or exceeded traditional NWP for many synoptic-scale variables. ECMWF operationalized AI ensemble systems in summer 2025. NOAA operationalized AI ensemble systems in Dec 2025.

The Question

Can AI-based ensemble forecasts improve prediction of aggregate wind power ramp events — where it matters most for system economic outcomes?

Background and Prior Work

AI Weather Prediction Breakthroughs

Lam et al. (2023, Science): GraphCast outperforms ECMWF HRES on 90% of 1,380 verification targets with 99.7% less compute

NOAA Project EAGLE (Dec 2025): Operationalized AIGEFS (GraphCast ensemble) and HGEFS alongside physics-based GEFS

Wind Power Ramp Forecasting

Hodge et al. (2014, WFIP): Ensemble NWP improves ramp forecasting at regional scale

Bossavy et al. (2015): Probabilistic ramp forecasting with autoregressive logit models

Draxl et al. (2015): Ensemble NWP economic value for grid operators demonstrated

No published evaluation of NOAA AI ensemble systems for aggregate wind power ramp prediction in an operational context

How AI Ensembles Differ from Physics Ensembles

System	Members	Initial Conditions	Propagation	Diversity Source
GEFS	31	Perturbed ICs (EnKF)	FV3 physics + stochastic physics	IC + physics perturbations
AIGEFS	31	Same GEFS perturbed ICs	GraphCast w/ distinct <i>calibrated</i> weight sets	IC + ML diversity
HGEFS	62	Combined GEFS + AIGEFS ICs	Grand ensemble (physics + ML)	IC + physics + ML diversity
PMLGEFS*	31	Same GEFS perturbed ICs	Graphcast w/ distinct <i>uncalibrated</i> weight sets	IC + ML diversity
ECMWF AIFS ENS	51	IFS ENS perturbed ICs	AIFS model (Anemoi framework)	IC perturbations
ECCC GEML	Det. only	GDPS analysis	Fine-tuned GraphCast	N/A (deterministic)

*Not currently operational

Study Design

Study Area: 140 wind farms across MN/IA (17,000 MW aggregate capacity)

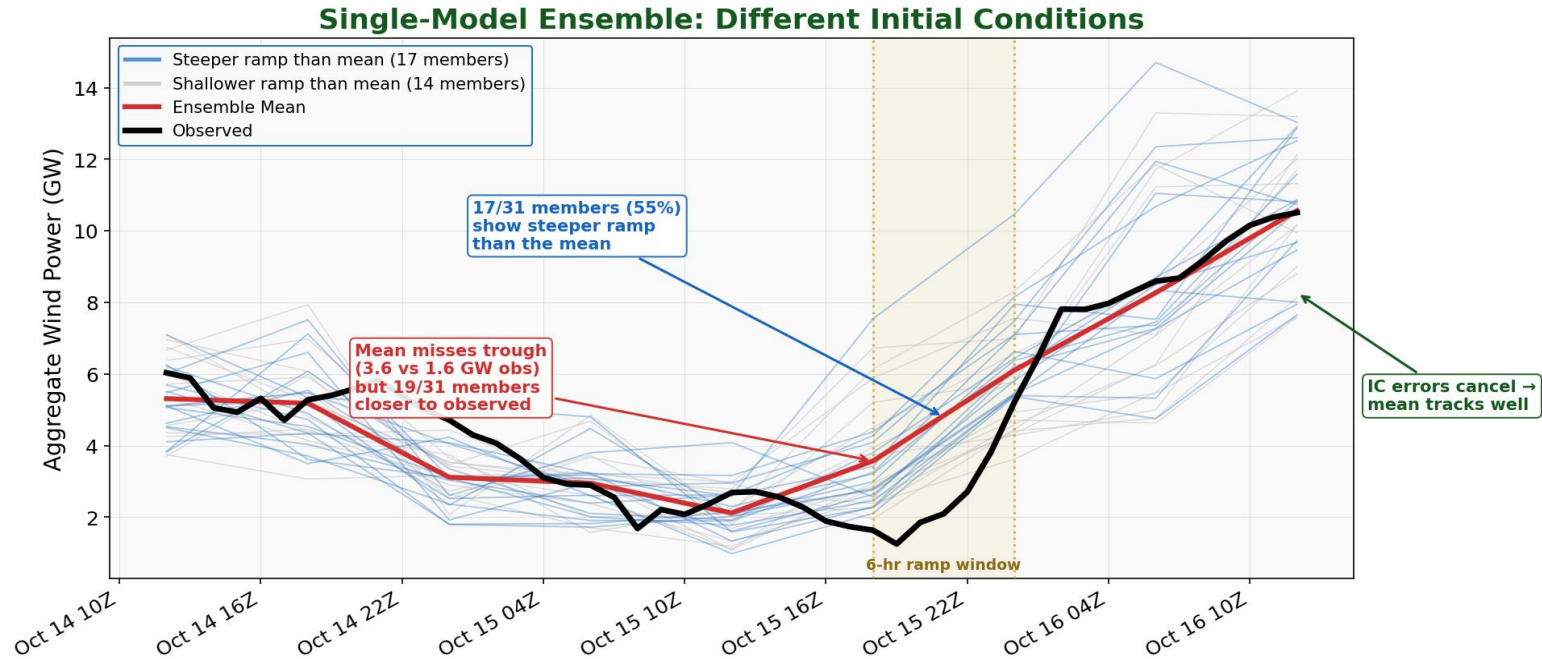
Period: October 1 – November 19, 2025 (50 days, autumn transition)

Ramp Definition: 6-hour rolling aggregate power change $\geq 15\%$ of portfolio capacity (2,520 MW), separated by up-ramp and down-ramp direction

Model Data: ensemble fixed-altitude wind and ambient condition predictions → power model conversions → portfolio aggregate

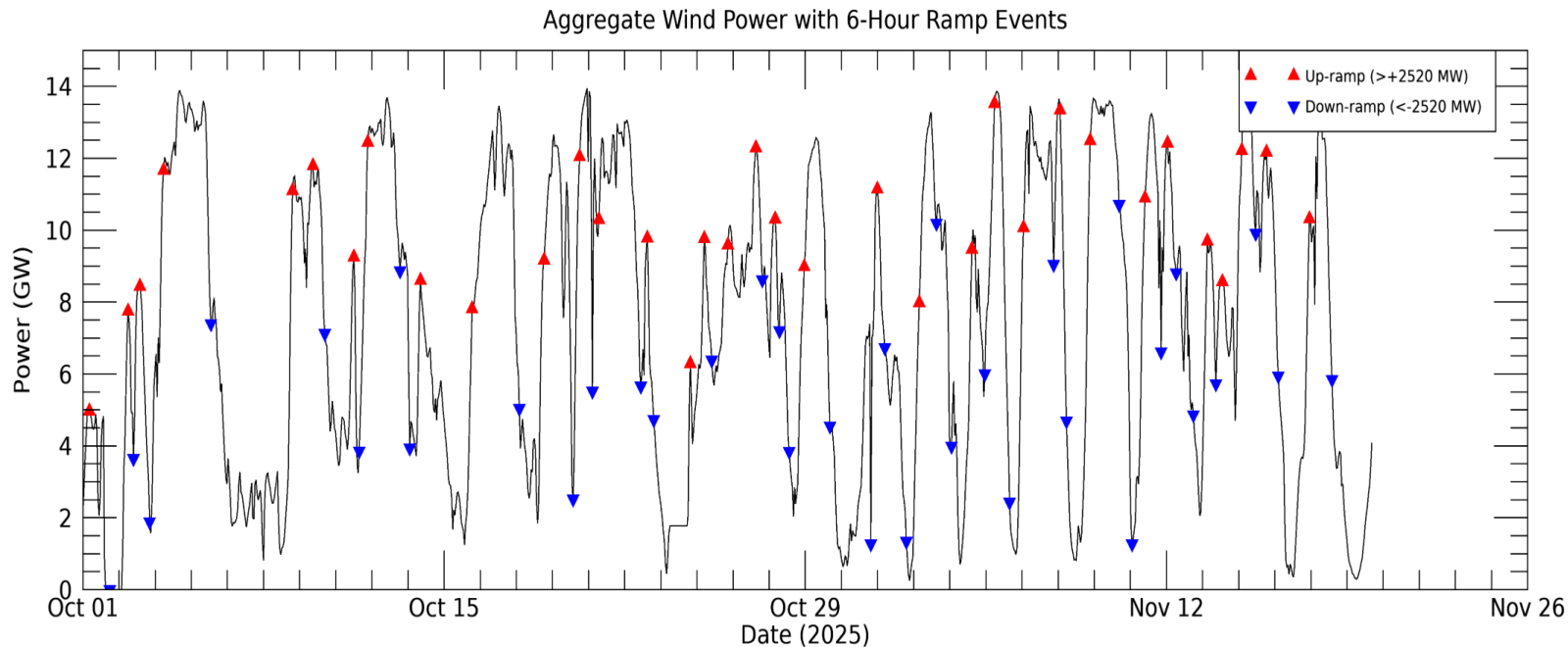
Calibration: Bias-Corrected Variance Inflation (BCVI) applied per model and lead time

Ensemble Systems Address Uncertainties



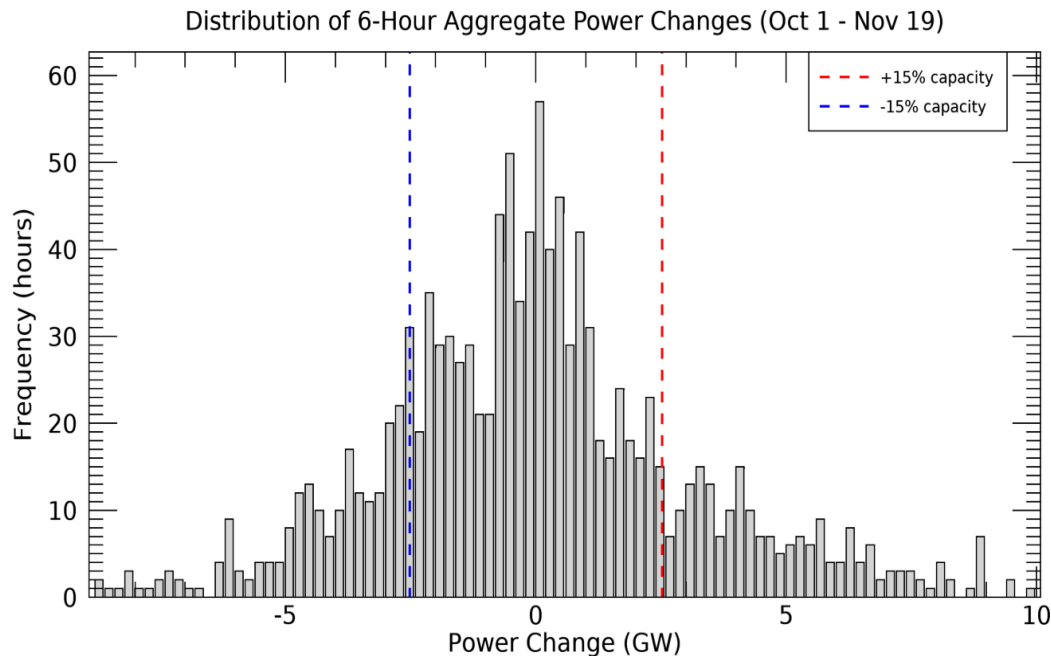
Oct 15–16, 2025 up-ramp (+3.6 GW). Before the ramp, the ensemble mean misses the trough — but 55% of members show a steeper ramp. After the event, IC errors cancel and the mean recovers. The ensemble spread provides an early probability signal the mean alone cannot.

Ramp Event Climatology



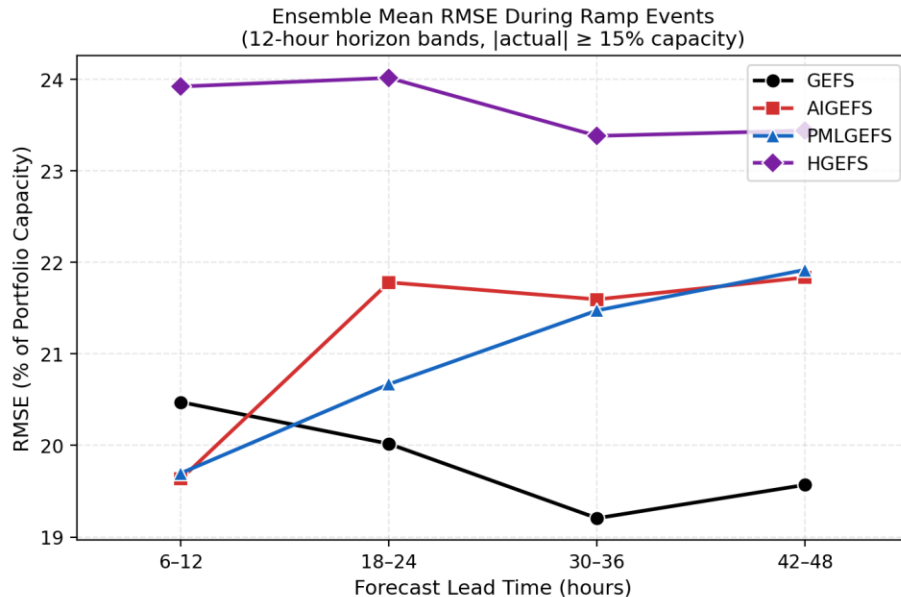
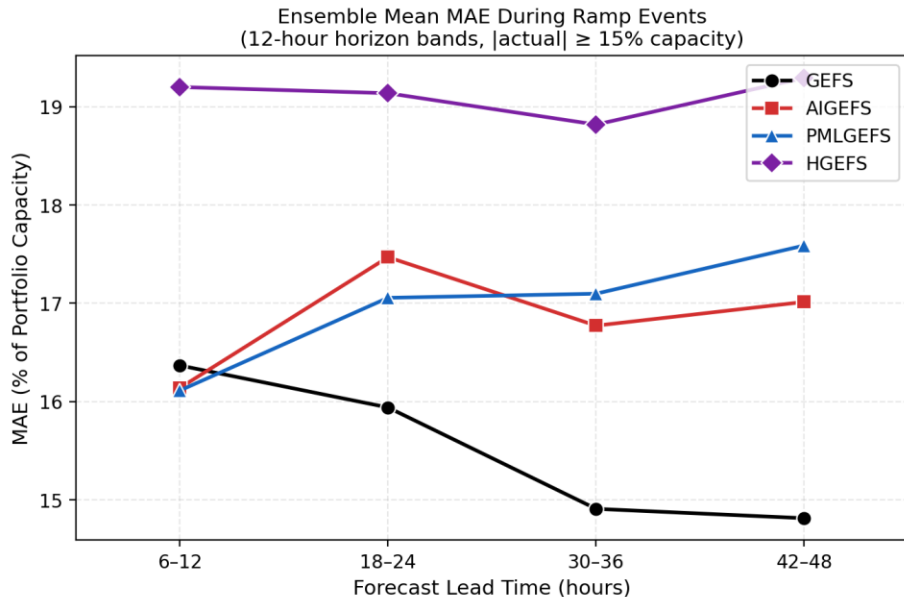
Approx. 100 ramp events detected over 50 days (mix of up-ramps and down-ramps).
Largest ramps often associated with frontal passages.

Ramp Event Climatology



Approx. 100 ramp events detected over 50 days (mix of up-ramps and down-ramps). Ramp activity correlates with frontal passages during autumn transition.

Ensemble Mean Performance: Ramp Events

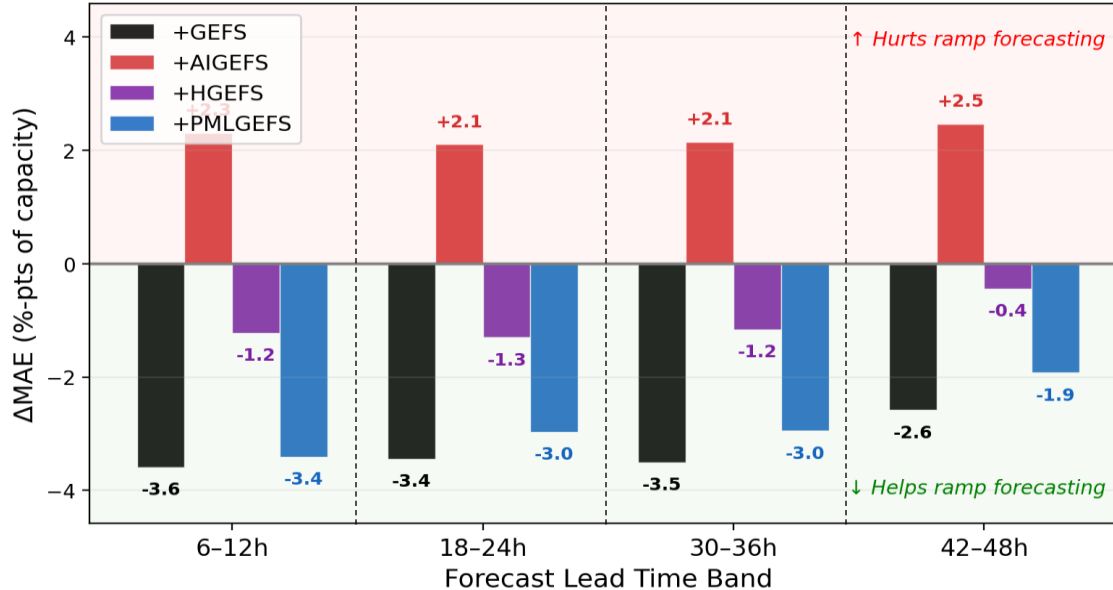


GEFS has the lowest ramp MAE, improving to ~15% at Day 2.

HGEFS is worst during ramps (~19%) despite largest ensemble (62 members) — AIGEFs members dilute ramp skill.

Additive Value: Blending with Deterministic NWP

Change in Ramp MAE: 30% Ensemble + 70% Det NWP vs. Det NWP Only
(Ramp events: $|\Delta P| \geq 15\%$ of portfolio capacity)

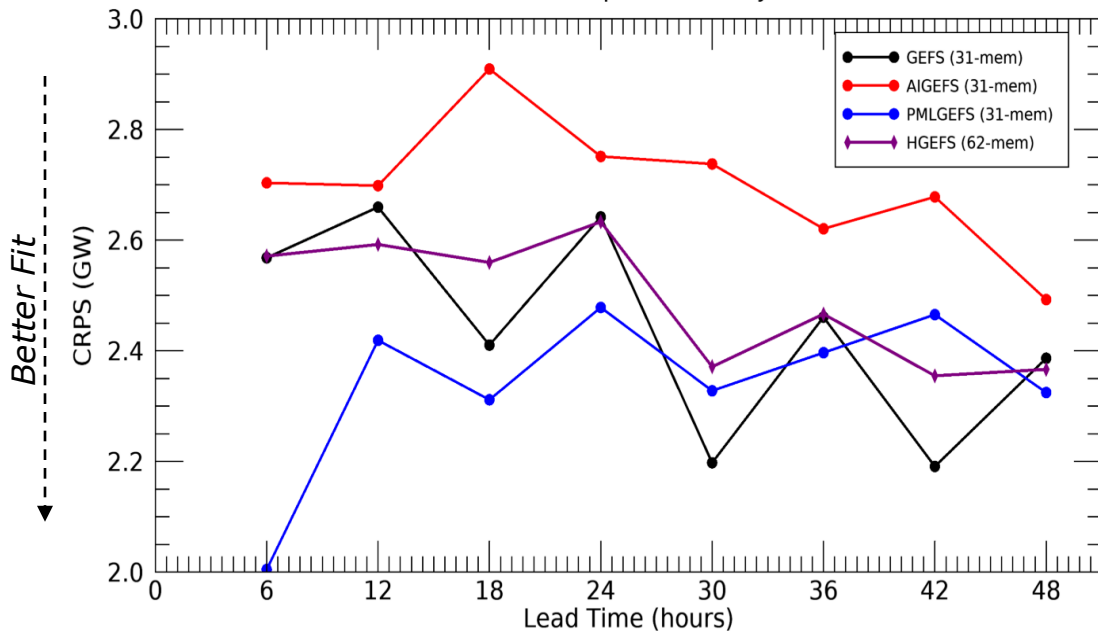


Blended at 30% weight with 12-model det. NWP blend:

- +GEFS** most consistent MAE reduction (avg -3.3% -pts)
- +HGEFS** slight improvement; diluted by AIGEFS members
- +PMLGEFS** competitive with GEFS (avg -2.8% -pts)
- +AIGEFS** degrades ramp prediction (avg $+2.3\%$ -pts)

Probabilistic Skill: CRPS All Ramps

CRPS on 6-Hour Ramp Forecasts by Lead Time



Best CRPS:

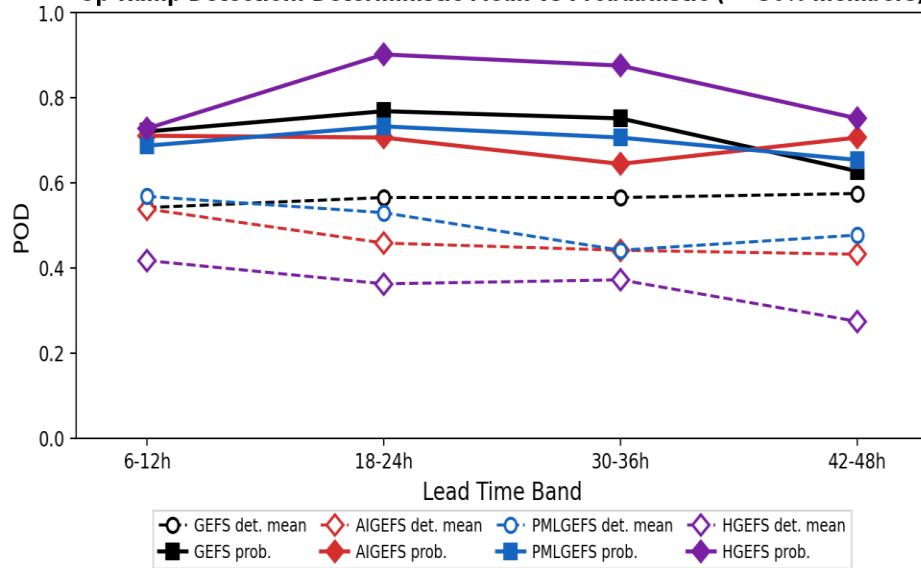
PMLGEFS has the lowest CRPS at most leads — well-calibrated probabilistic forecasts despite higher mean error.

AIGEFS has the highest CRPS.

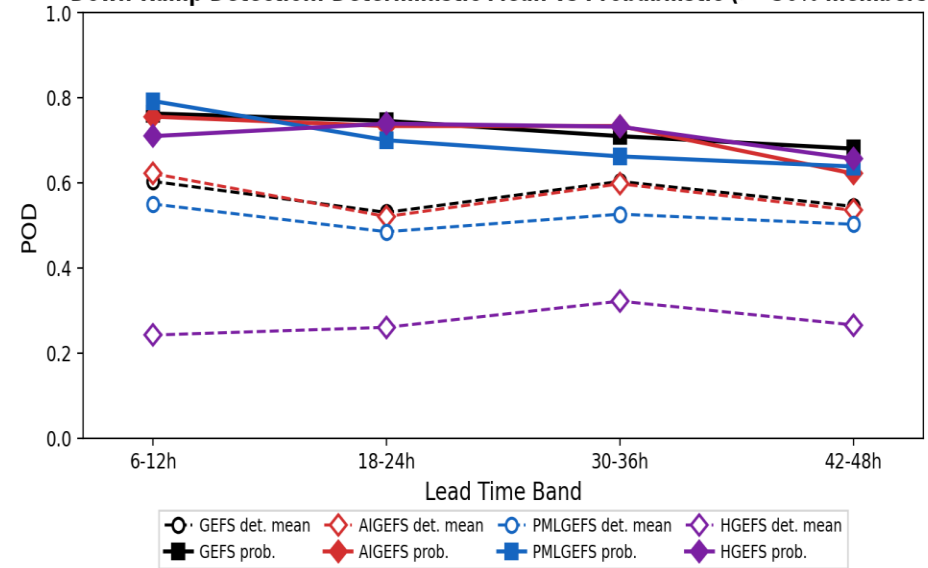
GEFS and **HGEFS** are intermediate.

Ramp Detection: Deterministic vs Probabilistic

Up-Ramp Detection: Deterministic Mean vs Probabilistic ($\geq 30\%$ members)



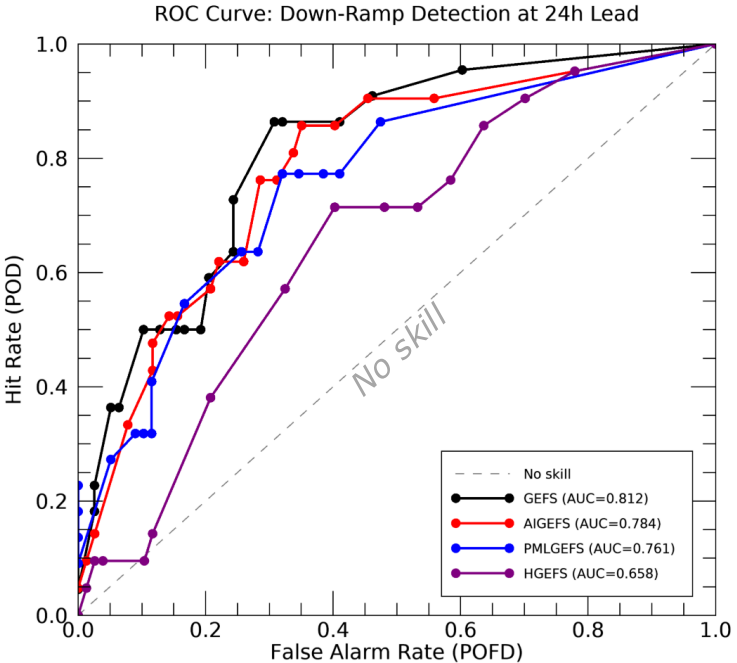
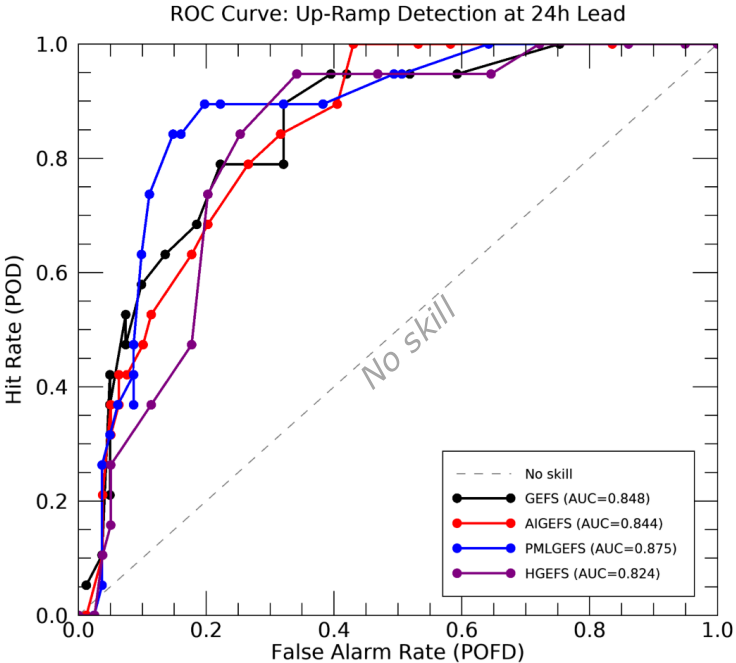
Down-Ramp Detection: Deterministic Mean vs Probabilistic ($\geq 30\%$ members)



Probabilistic detection ($\geq 30\%$ of members) improves POD for all models. **HGEFS** (62 mbrs) jumps from worst det. to best prob. up-ramp POD.

More members boost detection, but high POD alone does not guarantee forecast value.

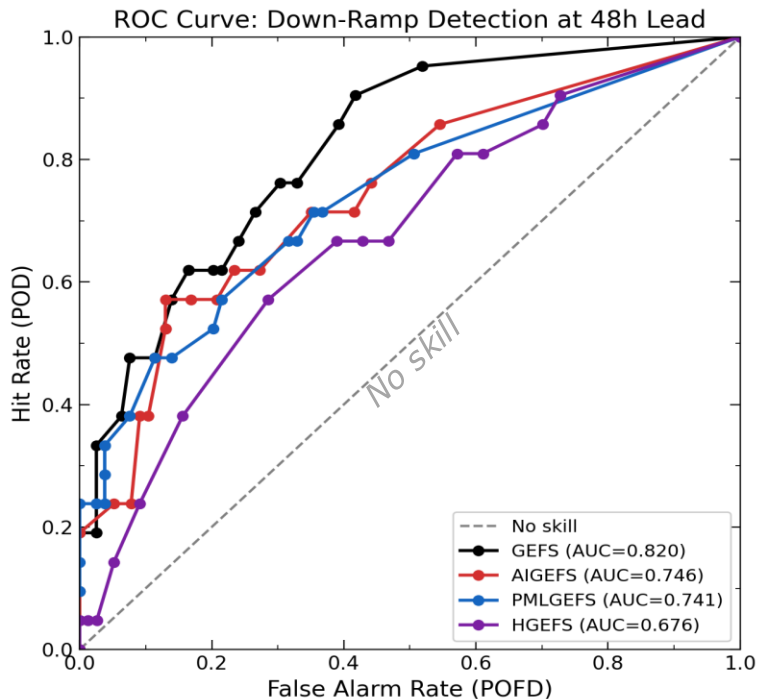
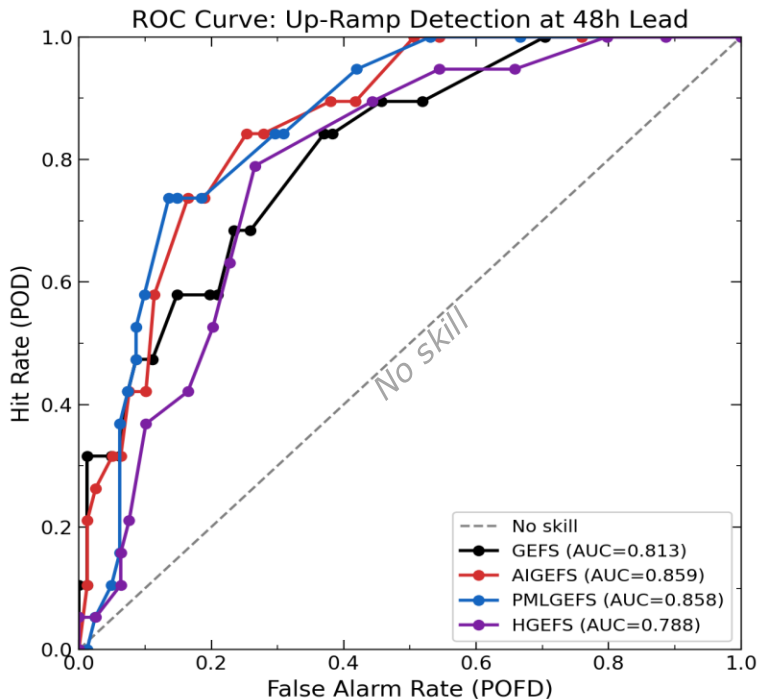
Discrimination: "ROC" Curves at 24-Hour Lead



Up-Ramps: PMLGEFS best AUC (0.875).
Down-Ramps: GEFS best (0.812).

HGEFS worst for both — combining degrades discrimination.

Discrimination: "ROC" Curves at 48-Hour Lead



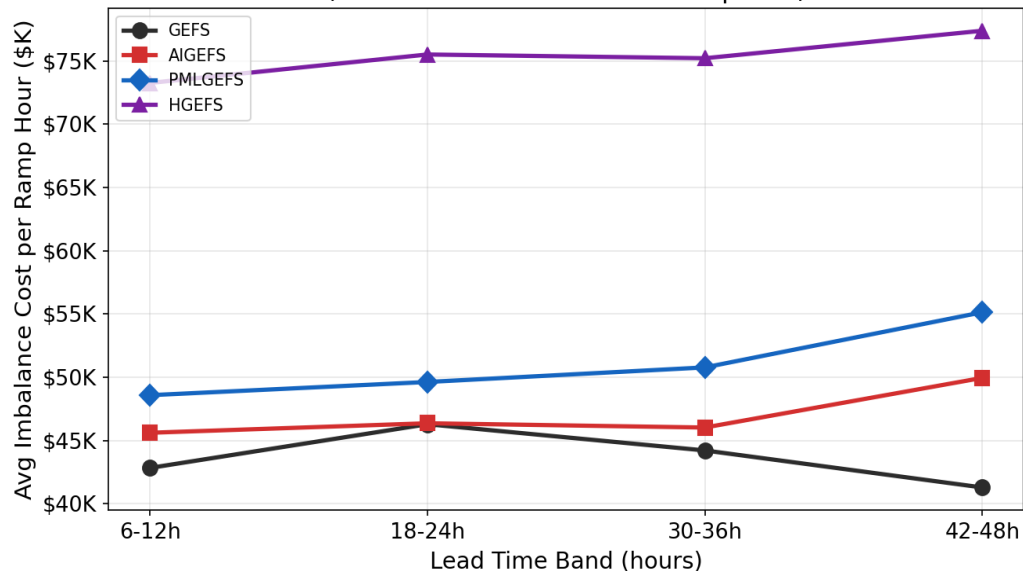
Up-Ramps: **AIGEFs** and **PMLGEFS** tied for best AUC (~0.86).

Down-Ramps: **GEFS** best (0.820).

HGEFS worst for both — combining degrades discrimination at Day 2 as well.

Economic Impact of Ramp Forecast Quality

Economic Exposure During Ramp Events
(Forecast Error × DA-RT LMP Spread)



DA-RT Price Volatility:

Ramp hours: \$25/MWh avg |DA-RT| spread
(52% above non-ramp)

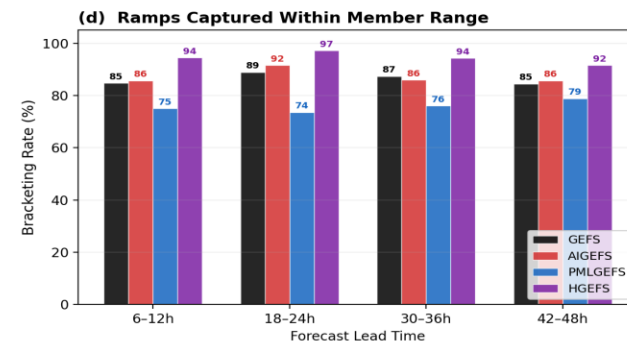
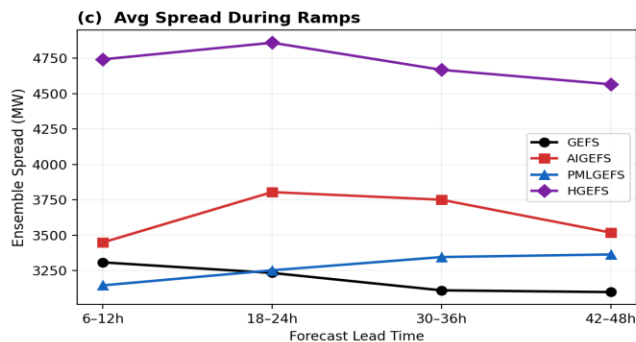
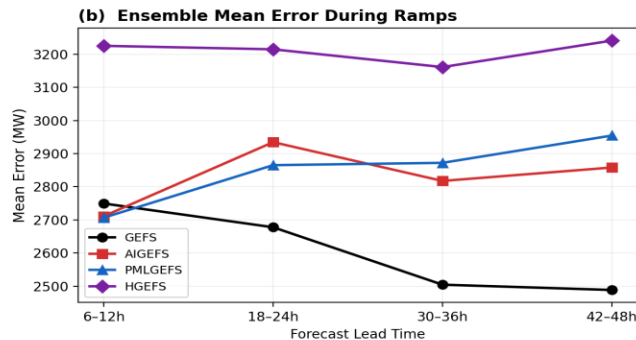
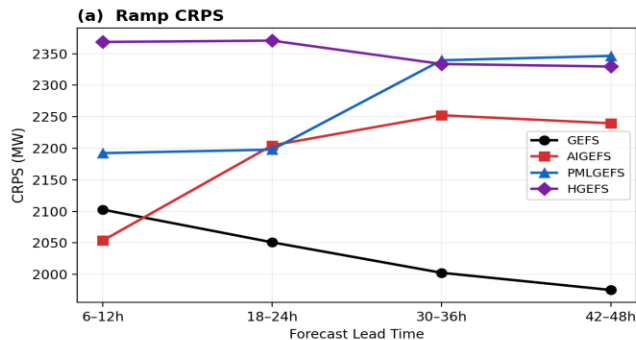
Portfolio Exposure:

~\$22.5M/yr imbalance exposure for this 17
GW portfolio

GEFS has lowest exposure overall.
HGeFS creates an expensive dilution effect.

Ensemble Ramp Performance Summary

Ensemble Performance During Ramp Events ($|\Delta P| \geq 15\%$ capacity)
Oct 1 - Nov 19, 2025 | ~16,800 MW Aggregate Wind Portfolio | IA/MN



Best ramp skill:

GEFS has lowest ramp CRPS and MAE at all leads. NWP diversity outperforms AI members for ramp forecasting.

Most events captured:

HGeFS brackets 92–97% of ramp events (62 members) but has worst MAE and CRPS — wide spread doesn't improve accuracy.

Implication:

GEFS provides best ramp accuracy and competitive event capture (85–89%). More members ≠ better ramp skill.

Key Findings

- 1. Ensemble means underpredict ramp magnitude across all model families**
Regression to the mean: ensemble means capture only 50–83% of actual ramp magnitude. This is not AI-specific — GEFS shows the same bias.
- 2. Ensemble construction matters more than member count**
HGEFS (62 members) has the worst ramp accuracy; naive model-family mixing dilutes the ramp signal. PMLGEFS (uncalibrated GC weights) outperforms AIGEFS (calibrated GC weights). Also different penalty functions: PMLGEFS minimized gridpoint error, while AIGEFS minimizes spectral harmonic error.
- 3. Detection skill \neq forecast skill**
AIGEFS and PMLGEFS tie for best up-ramp detection (AUC \sim 0.86), but only PMLGEFS improves deterministic NWP when blended (-2.8% MAE). AIGEFS degrades it ($+2.3\%$).
- 4. GEFS remains the benchmark for ramp accuracy and additive value**
Lowest CRPS and MAE at all leads, strongest blending improvement (-3.3% -pts MAE). NWP ensemble diversity still outperforms AI model diversity for ramp magnitude prediction.

Future Directions

Selective Member Weighting

Rather than equal-weight grand ensembles, use regime-dependent weighting to emphasize AI members when they show skill and suppress them when they add noise

Ramp-Optimized Calibration

Current BCVI calibration is optimized for overall performance. Ramp-specific calibration targeting tail events could unlock more AI ensemble value

Ensemble Subsetting for Event Types

PMLGEFS for up-ramps, GEFS for down-ramps — exploiting direction-dependent skill differences in an operational framework

Longer Evaluation Periods and Seasonal Dependence

This 50-day autumn study needs replication across seasons and geographies to establish robustness of the direction-dependent AI skill pattern

Thank You

Craig Collier

Chief Meteorologist, Head of Operations
Energy Forecasting Solutions

craig.collier@energyforecastingsolutions.com

Data: NOAA Project EAGLE (GEFS, AIGEFs, PMLGEFS, HGEFS)
Platform LMPs: Midwest Hub (public market data)

